甲型H1N1流感病毒HA基因的简单重复序列预测

吴 静[1],丁 勇[1*],刘成友[2]

([1]南京医科大学数学与计算机教研室,[2]生物医学工程系,江苏 南京 210029)

[摘 要] **目的**:探讨应用ARIMA模型对甲型H1N1流感病毒血凝素(hemagglutinin,HA)基因简单重复序列(simple sequence repeats,SSRs)的相对丰度和相对密度值进行预测的可行性,为防控流感流行制定措施提供科学依据。**方法**:应用Eviews 6.0软件对1970~2007年38条同源性相对较高的甲流H1N1流感病毒HA核苷酸序列中SSRs的相对丰度和相对密度进行拟合,建立时间序列模型,用模型对2008~2010年SSRs的相对丰度和相对密度进行预测,并用实际数据评估模型预测效果,进而预测2011年数据。**结果**:ARIMA模型较好地拟合了既往相对丰度和相对密度的实际序列,对2008~2010年的相对丰度和相对密度的预测也获得了较好的预测效果。**结论**:ARIMA模型能较好地模拟甲型H1N1流感病毒HA基因中SSRs的相对丰度和相对密度的变动趋势,可用于SSRs相对丰度和相对密度值的短期预测和动态分析。

[关键词] 时间序列;ARIMA模型;简单重复序列;甲型H1N1流感病毒;预测

[中图分类号] R183.3 [文献标志码] A [文章编号] 1007-4368(2013)01-042-06

doi:10.7655/NYDXBNS20130109

Forecast of SSRs in hemagglutinin sequences of influenza viruses A/H1N1

Wu Jing[1], Ding Yong[1*], Liu Chengyou[2]

([1]*Department of Mathematics and Computer*, [2]*Department of Biomedical Engineering*, *NJMU*, *Nanjing 210029*, *China*)

[**Abstract**] **Objective**: To explore the feasibility of using Autoregressive Integrated Moving Average (ARIMA) model to predict the relative abundance and relative density of simple sequence repeats (SSRs) in Hemagglutinin Sequences of influenza viruses A/H1N1, and to provide scientific basis for measures of preventing and controlling influenza pandemic. **Methods**: Eviews 6.0 software was utilized to construct the ARIMA model based on the relative abundance and relative density of SSRs in hemagglutinin (HA) sequences of influenza A with high homology from 1970 to 2007, and the constructed model was applied to predict the relative abundance and relative density from 2008 to 2010. The model was evaluated by actual data and then used to forecast the data of 2011. **Results**: The ARIMA model exactly fitted the relative abundance and relative density of the previous time series, and got a good predicting result on the data of 2008 to 2010. **Conclusion**: The ARIMA model can be used to make a short-term prediction and a dynamic analysis on the relative abundance and relative density of SSRs.

[**Key words**] time series; ARIMA model; simple sequence repeats; influenza viruses A/H1N1; prediction

[Acta Univ Med Nanjing, 2013, 33(1): 042-047]

流感是一种反复出现的传染病,在全球有高发病率和高病死率。每年造成全球约5亿人患病,25~50万人死亡。流感病毒(influenza virus)属正黏液病毒科,流感病毒属,包括甲、乙、丙3型,其中甲型抗原变异性最强,能感染人类和其他动物,侵袭所有年龄组人群,经常引起世界性大流行。甲型流感病毒根据其表面的血凝素(hemagglutinin,HA)和神经氨酸酶(neuraminidase,NA)基因的不同又可分成16个HA亚型(H1~H16)和9个NA亚型(N1~N9)。2009年2月,甲型流感病毒H1N1在墨西哥发现并开始在全球流行。据世界卫生组织统计,截至2010年2月,全球200多个国家和地区共发现40多万例确诊病例,其中已造成至少14 711人死亡。甲型流感病毒基因组含有8个RNA片段,其中第4区段所编码的

[基金项目] 南京医科大学基础医学院优势学科教师培养基金项目(JX10131801099)

*通信作者(Corresponding author),E-mail:yding@njmu.edu.cn

HA 是甲型流感病毒的主要抗原基因。HA 抗原具有免疫原性,能使人体产生保护性抗体,但其容易变异是流感流行的主要原因之一。研究发现,甲型流感病毒 HA 基因极易发生点突变,其机制涉及核苷酸序列中碱基的漂移和转换,最终导致编码蛋白的氨基酸序列改变[1]。

简单重复序列(simple sequence repeats,SSRs),也称微卫星序列(microsatellite)或短串联重复(short tandem repeat,STR),是一种以 1~6 个碱基为重复单元的 DNA 序列。随着研究的深入,人们发现 SSRs 在基因组上的分布并不是随机的,而且 SSRs 具有非常重要的生物功能。Usdin[2]认为 SSRs 与基因的重复、转录、失活和基因的稳定等方面都有着密切的关系[2]。其他研究表明,SSRs 具有很高的可变性、显著的多态性和侧翼序列的保守性,一般为共显性,其数量和种类还可能影响翻译活动的水平,并且与基因组的进化相关[3]。近年来,又有研究发现 SSRs 能够影响生物体中染色质的结构、基因活性的调控、DNA 的重组及错配修复系统,还可能与人类的癌症和遗传性紊乱相关,并在生物的进化过程中起着非常重要的作用[4]。

在 SSRs 分布规律的研究中,相对丰度和相对密度是两个重要的指标。所谓相对丰度,是指序列单位长度中含有的各碱基 SSRs 的数量;相对密度是指各碱基 SSRs 的长度总和占整个序列长度的比例。在甲型流感病毒 HA 基因组中,存在着大量的简单重复序列。在生物进化的过程中,不同甲型流感病毒株的简单重复序列由于突变偏好,或面临不同的选择压力,导致相对丰度和相对密度上的差异。因此,SSRs 的相对丰度和相对密度的状况蕴藏着丰富的关于基因序列结构、功能和进化的信息,SSRs 的相对丰度和相对密度的研究对于探讨甲型流感病毒的变异性具有重大的意义。本研究采用 ARIMA 模型对 1970~2007 年 38 条同源性相对较高的甲型流感 HA 核苷酸序列中 SSRs 的相对丰度和相对密度进行拟合,并对 2008~2011 年的相对丰度和相对密度进行预测与检验,为制定防控流感以及季节性流感流行的措施提供依据。

1　资料和方法

1.1　资料

数据资料来源于 NCBI 网站,网址为:http://www.ncbi.nlm.nih.gov/,选择同源性相对较高的 41 条序列,表 1 列出了每一条 HA 基因序列的 GenBank 登记号、基因序列长度和来自的国家或地区。

选择简单重复序列的基本原理是前一个简单重复序列被单一碱基隔开后即开始查找下一个简单重复序列。我们用在线查找软件 IMEx[5]来查找 HA 片段全序列中存在的所有一、二、三、四、五、六型 SSRs,生成原始的数据统计结果。在筛选过程中,考虑到甲型流感病毒 HA 序列基因组较小的实际情况,查找时选择"完全重复"类型。此外,为了避免因碱基随机组合形成的重复序列,设置查找条件[6]:单碱基和二碱基 SSRs 重复次数≥3,其余 SSRs 重复次数≥2。把查找到的结果写入 Excel 2007 中,算出每条序列的每1 000 bp 中各型 SSRs 的相对丰度值和各型 SSR 长度总和的相对密度值(表 1)。例如:S1 的编号为 CY022444,基因组总长度是 2 309 bp,单碱基 SSRs 为 116 个(其中重复 3 次 74 个,重复 4次 37 个,重复 5 次 4 个,重复 6 次 1 个),其碱基长度总和为 396 bp,则单碱基 SSRs 相对丰度值为 (116/2 309) × 1 000≈50.24,相对密度值为 (396/2 309) × 1 000≈171.50。本研究采用 1970~2007 年的数据建立 ARIMA 模型,预测 2008~2011 年的数据,并用 2008~2010 年的实际数据验证模型预测效果。

1.2　*方法*

20 世纪 70 年代,美国统计学家 Box 和英国统计学家 Jenkins 提出了一整套关于时间序列分析、预测和控制的方法,被称为 Box-Jenkins 建模方法。求和自回归移动平均(autoregressive integrated moving average,ARIMA) 模型是其中重要而基本的模型之一,该方法的基本思想是将预测对象随时间推移而形成的数据序列视为一个随机序列,即除去个别的因偶然原因引起的观测值外,时间序列就是对随机过程进行观测所取得的一组离散观测。这组随机变量所具有的依存关系或自相关性表现了预测对象发展的延续性,而这种自相关性一旦被相应的数学模型描述出来,就可以从时间序列的过去值及现在值预测未来值。

ARIMA 是自回归 AR、求和 I、移动平均 MA 三个模型的混合,是指序列 X_t 的 d 阶差分具有 ARMA(p,q)模型,记作 ARIMA(p,d,q)。该模型的数学形式为:

$$\Phi(B)\nabla^d X_t=\Theta(B)\varepsilon_t$$

其中,t 代表时间,X_t 表示响应序列,B 是后移算子,$\nabla=1-B$,p、d、q 分别表示自回归阶数、差分阶数和移动平均阶数,$\Phi(B)=1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p$ 为平稳可逆 ARMA(p,q)模型的自回归系数多项式,

表 1 HA 核苷酸序列来自的地区和序列长度,以及一、二型 SSRs 的相对丰度及相对密度值

Table 1 List of analyzed HA sequences, their attributed regions and length, relative abundance and relative density of mono, di repeats in HA sequences

编号	GenBank 登记号	序列长度(bp)	来源国家或地区	时间(年份)	相对丰度值	相对密度值
S1	CY022444	2309	Wisconsin	1970	50.238/3.465	171.503/21.654
S2	CY022420	2309	Wisconsin	1971	51.104/3.898	172.802/24.253
S3	CY022419	2299	Wisconsin	1971	55.241/0.870	189.648/5.220
S4	EU139826	1701	Iowa	1973	58.201/2.352	195.767/14.109
S5	X57491	1074	HongKong	1974	53.073/1.862	178.771/11.173
S6	CY024940	2297	Minnesota	1975	50.065/3.483	169.351/21.768
S7	CY036806	2295	Minnesota	1976	50.980/3.922	172.113/24.401
S8	CY032936	2289	Tennessee	1977	49.367/3.932	167.759/24.465
S9	CY028422	1522	Illinois	1978	53.219/5.256	172.142/31.537
S10	CY019746	2293	Memphis	1979	50.153/2.181	172.263/13.956
S11	CY023011	988	Auckland	1980	45.547/1.012	158.907/6.073
S12	CY022985	2277	Ontario	1981	51.383/3.513	172.156/21.959
S13	CY010371	2333	Baylor	1982	51.865/2.572	177.025/17.145
S14	CY037965	2280	Belgium	1983	44.737/2.193	153.509/13.158
S15	CY037972	2280	France	1984	46.491/1.754	159.649/10.526
S16	CY037980	2280	France	1985	46.053/1.754	158.333/10.526
S17	CY028795	2309	Iowa	1986	51.537/3.032	170.637/19.056
S18	CY022969	2301	Iowa	1987	52.151/3.911	174.707/25.206
S19	CY039932	2292	Indiana	1988	52.360/3.490	174.084/21.815
S20	CY037988	2280	Belgium	1989	46.930/1.316	160.965/7.895
S21	CY035077	2289	Memphis	1990	51.349/3.046	169.713/19.147
S22	CY027162	2283	Iowa	1991	50.810/3.066	167.762/19.273
S23	CY038004	2280	England	1992	43.860/1.754	150.877/10.526
S24	CY037903	2280	England	1993	44.298/1.754	152.193/10.526
S25	CY037911	2280	England	1994	45.175/1.754	155.263/10.526
S26	CY037919	2280	England	1995	45.175/1.754	155.702/10.526
S27	CY037934	2274	England	1996	54.969/0.880	182.498/5.277
S28	CY037941	2280	England	1997	44.298/1.754	152.632/11.404
S29	CY037949	2280	England	1998	44.737/1.754	154.386/11.404
S30	CY037957	2279	Scotland	1999	48.267/2.194	164.107/13.164
S31	FJ752499	2280	Changhua	2000	46.491/2.632	156.579/16.667
S32	GQ229361	2250	Hong Kong	2001	48.444/2.667	165.333/16.000
S33	HM125972	2280	MN	2002	50.877/2.632	171.053/15.789
S34	CY010579	2309	Spain	2003	46.773/2.165	156.778/13.859
S35	CY010587	2309	Spain	2004	48.073/2.165	161.975/13.859
S36	FJ638303	2285	NC	2005	49.891/3.501	166.302/21.007
S37	CY035427	2251	Minnesota	2006	48.867/2.665	163.483/15.993
S38	CY035445	2252	Illinois	2007	47.069/2.664	159.414/15.986
S39	CY042313	2342	Illinois	2008	49.530/2.562	166.951/15.371
S40	CY061826	2345	Hong Kong	2009	49.032/2.581	167.742/15.484
S41	CY061803	2323	Hong Kong	2010	53.810/2.580	179.930/16.360

$\Theta(B)=1-\theta_1B-\theta_2B^2-\cdots-\theta_qB^q$ 为平稳可逆 ARMA(p,q)模型的移动平均系数多项式。ε_t 为零均值白噪声序列,代表独立扰动或随机误差。ARIMA(p,d,q)模型拟合数据的过程, 实质上是先对观测数据进行 d 次差分处理,然后再拟合 ARMA(p,q)模型。

2 结 果

2.1 序列的平稳化

图 1A 是 1970~2007 年甲型流感病毒 HA 核苷酸序列中二碱基 SSRs 相对丰度的时间序列图,样本

容量为 38。由图 1A 可看出该时间序列的波动较大,考虑对原始序列做一阶差分,从新序列的自相关图和偏相关图(图 1B)可以看到,数据的平稳性得到很大改进,选择不含常数项和趋势项的 ADF 检验,检验结果也显示数据呈现平稳性。

2.2　模型的识别

从自相关图和偏相关图(图 1B)可以看出,自相关系数和偏相关系数均呈拖尾现象,由上可初步判断模型为 ARIMA(2,1,1)。为保证模型预测的效果,p 和 q 的最终确定还要从低开始试探,如 ARIMA(0,1,1)、ARIMA(1,1,0)、ARIMA(1,1,1)、ARIMA(2,1,0)、ARIMA(2,1,1),根据模型的拟合优度、残差情况以及系数间的相关性进行筛选,定出合适的模型。

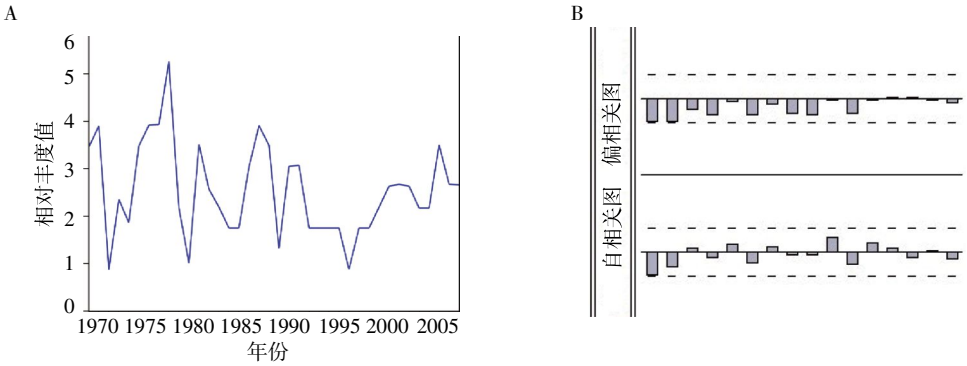

图 1　实际二碱基 SSRs 相对丰度序列图(A)和差分后序列的自相关图偏相关图(B)

Figure 1　Time series of relative abundance of di repeats in HA sequences(A) and autocorrelation (below) and partial correlation (above) of series after one time differenced(B)

2.3　参数估计及检验

经 Eviews 6.0 计算,2 个备选模型的参数具有统计学意义,结果如下(表 2)。进一步参照拟合优度及 AIC 准则,确定选用模型 ARIMA(0,1,1)。

2.4　模型的诊断

在 Eviews 6.0 中选择残差检验的 Q 统计相关图检验,结果如下(表 3)。根据各滞后期 Q 统计量的 P 值,检验结果不能拒绝残差不相关的零假设,即模型 ARIMA(0,1,1)的残差序列是白噪声序列,从而提示所选模型恰当,用于预测是合适的。

表 2　备选模型的参数估计结果

Table 2　Parameter estimates of alternative ARIMA models

参数	ARIMA(0,1,1)			ARIMA(2,1,0)		
	系数	t 值	P 值	系数	t 值	P 值
AR(1)	–	–	–	-0.374	-2.529	0.016
AR(2)	–	–	–	-0.315	-2.119	0.042
MA(1)	-0.993	-20.919	<0.001	–	–	–
R^2 值		0.306			0.201	
AIC		2.854			2.907	

表 3　残差序列的自相关检验

Table 3　Autocorrelation check of residuals

滞后阶数	Q 统计量	P 值	自相关函数值
6	3.131	0.680	0.238,0.004,0.017,-0.028,0.004,-0.133
12	6.064	0.869	-0.065,-0.132,-0.046,0.146,-0.011,0.102
18	10.419	0.885	0.049,-0.054,-0.052,-0.098,-0.034,-0.202
24	15.552	0.874	-0.098,-0.099,-0.132,-0.047, 0.014, 0.124

2.5　模型的预测

用 ARIMA(0,1,1)模型预测 2008~2011 年甲型流感 HA 核苷酸序列中二碱基 SSR 相对丰度,并结合 2008~2010 年实际数据进行预测精度的验证,结果如下(表 4)。由表 4 数据可知预测精度基本上达到 98%以上,说明 ARIMA(0,1,1)模型对甲型流感病毒 HA 核苷酸序列中二碱基 SSR 相对丰度的预测是非常有效的。

2.6　其他指标预测结果

表 5~7 用同样的方法对甲型流感病毒 HA 核苷

表 4 二碱基 SSRs 相对丰度的实际值与预测值

Table 4 Actual values and predictive values of relative abundance of di repeats

年份	二碱基 SSR 相对丰度			
	实际值	预测值	绝对误差	相对误差
2008	2.562	2.609	0.047	0.018 3
2009	2.581	2.596	0.015	0.005 8
2010	2.583	2.586	0.003	0.001 2
2011	–	2.578	–	–

表 5 单碱基 SSRs 相对丰度的实际值与预测值

Table 5 Actual values and predictive values of relative abundance of mono repeats

年份	单碱基 SSR 相对丰度			
	实际值	预测值	绝对误差	相对误差
2008	49.530	48.456	−1.074	−0.021 7
2009	49.032	48.915	−0.117	−0.002 4
2010	53.810	49.068	−4.742	−0.088 1
2011	–	49.112	–	–

表 6 单碱基 SSRs 相对密度的实际值与预测值

Table 6 Actual values and predictive values of relative density of mono repeats

年份	单碱基 SSR 相对密度			
	实际值	预测值	绝对误差	相对误差
2008	166.951	163.695	−3.256	−0.019 5
2009	167.742	166.474	−1.268	−0.007 6
2010	179.940	166.474	−13.466	−0.074 8
2011	–	166.474	–	–

表 7 二碱基 SSRs 相对密度的实际值与预测值

Table 7 Actual values and predictive values of relative density of di repeats

年份	二碱基 SSR 相对密度			
	实际值	预测值	绝对误差	相对误差
2008	15.371	15.939	0.568	0.036 9
2009	15.484	15.891	0.407	0.026 3
2010	16.358	15.856	−0.502	−0.030 7
2011	–	15.827	–	–

酸序列中单碱基 SSRs 相对丰度、单碱基 SSRs 相对密度以及二碱基 SSRs 相对密度做预测的结果，可见 ARIMA 模型较好地拟合了历史数据，对以上各指标的预测精度都较高。

3 讨 论

甲型流感病毒为常见流感病毒，极易发生变异，经常引起世界性大流行，造成多人死亡，已引起世界各国卫生组织的高度重视。甲型流感病毒 HA 基因极易发生点突变，对其简单重复序列的相对丰度和相对密度建立时间序列模型，可以较好地对其分布规律进行预测和监测，对于探讨甲型流感病毒的变异性具有重大的意义，从而为防控流感流行制定措施提供科学依据。传统的时间序列模型均假设各变量之间是一种线性关系，因而实际预测时会使预测值呈不断上升或下降的趋势，而不能按照原有实际情况拟合模型，效果往往不佳。ARIMA 模型拟合可以综合考虑序列演变的趋势、周期变化和随机干扰因素，借助模型参数的变化对数据进行量化表达，是一种精度较高的短期预测模型。在医学领域，近年来该模型已成功地应用于 DNA 碱基预测、传染病预测等方面[7-13]，但在 SSRs 方面的研究尚未见报道。ARIMA 模型的应用前提是时间序列的平稳性，本研究注意到这点，对不平稳的序列进行差分预处理，使之达到平稳，再建立 ARIMA 模型对甲型流感 HA 核苷酸序列中单碱基、二碱基 SSRs 相对丰度和相对密度值进行时间序列分析，因而模型较好地拟合了历史数据，经实际数据验证，预测精度较高。

由于时间序列预测是按一定时序的规律进行

的,其前提是数据在一定时间内保持相对稳定,如果事情原有的趋势发生了很大的改变,会出现预测值与实际值明显不符的情况,因此单次分析建立的模型不能作为永久不变的预测工具,若需进一步预测,就需要积累新的数据对模型进行修正,甚至重新拟合。此外,本研究的样本量不是很大,下一步研究可考虑增加样本量,将季节因素纳入模型,建立更能反映实际情况的预测模型,这需要我们对实际资料的特点有深刻的理解,并不断积累经验。

[参考文献]

[1] Lindstrom SE,Cox NJ,Klimov A. Genetic analysis of human H2N2 and early H3N3 influenza viruses 1957-1972: evidence for genetic divergence and multiple reassortment events[J]. J Virol,2004,328(1):101-119

[2] Usdin K. The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases[J]. Genome Res,2008,18(7):1011-1019

[3] Kashi Y,King D,Soller M. Simple sequence repeats as a source of quantitative genetic variation[J]. Trends Genet,1997,13(2):74-78

[4] Li YC,Korol AB,Fahima T,et al. Microsatellites:genomic distribution,putative functions and mutational mechanisms:a review[J]. Mol Ecol,2002,11(12):2453-2465

[5] Mudunuri SB,Nagarajaram HA. IMEx:imperfect microsatellite extractor[J]. Genome Analysis,2007,23(10):1181-1187

[6] 彭　军,谭钟扬,蔡立军,等. 甲型流感病毒 HA 基因的简单重复序列分布分析[J]. 生物信息学,2011,9(1):54-59

[7] 刘　娟, 高　洁. 甲型 H1N1 流感病毒 DNA 序列碱基的预测[J]. 生物信息学,2011,9(3):259-262

[8] 杨　娟,戚建江,沈　毅. ARIMA 模型在杭州市中小学生咳嗽症状监测中的应用[J]. 生物数学学报,2011,26(3):563-568

[9] 彭志行,鲍昌俊,赵　杨,等. ARIMA 乘积季节模型及其在传染病发病预测中的应用 [J]. 数理统计与管理,2008,27(2):362-368

[10] 张彦琦,唐贵立,王文昌,等. ARIMA 模型及其在肺结核预测中的应用[J]. 现代预防医学,2008,35(9):1608-1610

[11] 吴孟泉,赵　凯. 基于 ARIMA 模型的 2009 年山东省手足口病疫情分析及预测[J]. 鲁东大学学报:自然科学版,2011,27(1):71-75

[12] 丁晓艳,彭志行,陶　红,等. 运用时间序列模型对麻疹流行趋势的预测与分析[J]. 南京医科大学学报:自然科学版,2011,31(8):1200-1202

[13] 胡建利,祖荣强,彭志行,等. 江苏省戊型肝炎发病趋势的时间序列模型应用[J]. 南京医科大学学报:自然科学版,2011,31(12):1874-1878

[14] 方积乾,陆　盈,张晋昕,等. 现代医学统计学(时间系统分析方法及其医学应用) [M]. 北京:人民卫生出版社,2002:219-269

[15] 攸　频,张晓峒. Eviews 6 实用教程[M]. 北京:中国财政经济出版社,2008:117-150

[收稿日期] 2012-08-20